DOCUMENT RESUME

ED 210 621 CS 006 327

AUTHOR Brittain, Mary M.; Brittain, Clay V.
TITLE Domain-Referenced Testing of Reading Achievement.
PUB DATE Mar 81
NOTE 16p.; Paper presented at the Meeting of the Eastern Educational Research Association (Philadelphia, PA, March 12-14, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Reading Achievement; *Reading Tests; Test Construction; *Test Format; *Testing Problems; Test Reliability; *Test Theory; Test Validity
IDENTIFIERS *Domain Referenced Tests; Skills Monitoring System

ABSTRACT

A behavioral domain is well-defined when it is clear to both test developers and test users which categories of performance should or should not be considered for potential test items. Only those tests that are keyed to well-defined domains meet the definition of criterion-referenced tests. The greatest proliferation of criterion-referenced tests in reading has been in the skills monitoring systems or reading management systems. The program-dependent type of test is interlocked with a single published instructional program such as a basal reading series, which is, in effect, part of the program. The second type, program independent, is not integrated with any particular published reading series, but may include charts that cross-reference different basal reading series by page number. A notable weakness, however, is the paucity of information reported on the reliability and validity of the tests used in these systems. There are no published domain-referenced tests of reading behavior. There have, however, been a number of research efforts to develop such tests. These efforts have called attention to the difficulty and amount of time and effort involved in defining domains. What is needed are tests that fall between the so-called criterion-referenced tests and the domain-referenced tests--tests that are feasible in terms of the time and effort required to develop them, but which provide a sound informational base for instructional decisions. (HOD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED210621

# Domain-Referenced Testing of Reading Achievement

Mary M. Brittain
Virginia Commonwealth University
Richmond, Virginia

Clay V. Brittain
Army Training Support Center
Ft. Eustis, Virginia

Eastern Educational Research Association
Philadelphia, Pennsylvania
March 12-14, 1981

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Mary M. Brittain
Clay V. Brittain
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CS206327

## Introduction

In a paper presented at the 1969 Invitational Conference sponsored by Educational Testing Service, Guttman (1969) made use of a cartoon to portray what he considered to be the predicament of achievement testing. The cartoon pictured achievement testing as a cart with two highly disparate wheels: a very large wheel labelled "score theory" and a very small wheel labelled "content theory." The present symposium may reflect in microcosm that this continues to some extent to be the situation in which achievement testing finds itself. But over the past decade, there has been extensive work devoted to strengthening the content wheel and thus to correcting the imbalance portrayed by Guttman. The efforts have proceeded under a variety of labels: criterion-referenced testing, domain-referenced testing, objective-referenced testing, mastery testing, and so forth. That the effort has not been entirely systematic is suggested by this plethora of labels and by the variety of definitions assigned to the same term. According to Berk (1980), for example, more than 50 definitions of criterion-referenced testing (CRT) have appeared in the literature.

### Frame work for viewing CRT

A number of different writers have attempted to reduce this confusion by more coherently ordering the emerging trends in achievement testing. In a recent article, Nitko (1980) proposed a taxonomy which initially differentiates achievement tests on the basis of the definition of the behavioral domain to which the tests are referenced. These domains are characterized as either well-defined, ill-defined, or basically undefined. According to Nitko, a behavioral domain is well-defined when it is clear to both test developers and test users which categories of performance should or should not be considered as potential test items.

Only those tests which are referenced to well-defined domains meet the definition of criterion-referenced tests. But criterion-referenced tests may be further divided into those which are referenced to ordered domains and those which are referenced to well-defined but unordered domains. Tests of the first type are further subdivided in terms of the basis on which the domain is ordered; e.g. estimated social or esthetic quality of performance, complexity or difficulty level of subject matter. Tests referenced to unordered domains are subdivided in terms of the basis for delineating the domain; e.g. diagnostic categories of performance. Altogether, Nitko identified some 15 varieties of criterion-referenced tests and cited examples of each type. Interestingly, some of the examples cited by Nitko have been around for a long time: Rev. Fisher's Scale books (1964), Ayres' Spelling Scale (1915), Starch's English Vocabulary Test (1916).

Berk (1980) has proposed a conceptually simpler scheme for categorizing criterion-referenced tests. He argues that two major conceptualizations of criterion-referenced testing have emerged in the literature: domain-referenced testing and mastery testing. These two kinds of criterion-referenced testing differ in their approaches to content domain specification (instructional objectives versus other strategies such as item transformations, mapping sentences, algorithms), item construction (generation rules versus tradition rules), item domain specification, item analysis, item selection (random versus nonrandom), parallel forms assumptions (randomly parallel versus classically parallel), the use of cutting scores, methods of assessing validity and reliability, and score interpretation (reference score to domain or reference score to cutting score).

Since our interest here is mainly in tests of reading, a framework proposed by Waldrop et al (1978) for analyzing reading tests seems particularly relevant. However, it should be noted that this framework appears equally applicable to tests in other content areas.

The point of departure in Waldrop's framework is the distinction made by Buros (1977) between testing for differentiation and testing for measurement. Differentiation places examinees in categories for selection decisions where access is limited. Thus, rank-ordering of examinees is important. Measurement, on the other hand, estimates the quantity of a characteristic and provides information for decisions about individuals where access is, at least in principle, open to all. The framework described by Waldrop and his associates places at one end of a continuum standardized achievement tests used for differentiation and at the other end domain-referenced tests used for measurement. The test-use dimension is paralleled by four other dimensions having to do with methods of item generation, item revision, methods of assessing test precision (reliability), and methods of assessing validity. In other words, the planned usage of a test implies certain methods of item generation and revision and certain methods of assessing reliability and validity.

Waldrop et al believe it useful in assessing tests to map them on the dimensions of their framework. A "good" test profile would be a reasonably straight line. According to the proponents of the framework, a straight line running down the left side or the right side of the scales would indicate a consistent inference system. An important implication here is that the framework is not only descriptive but is also prescriptive and is useful in evaluating achievement tests. The profiles of the standardized reading achievement tests used in the Anchor Test Study (Jaeger, 1973) were described as a consistent

profile on the left-hand portion of the scales. Waldrop et al found no reading achievement tests which could be profiled down the right-hand portion of the scales; i.e. no domain-referenced tests of reading achievement.

Now against this background we wish to comment on commercially available reading tests which have been identified by their publishers as criterion-referenced tests and about research/exploratory efforts to develop domain-referenced tests in reading.

Criterion-Referenced Tests in Reading: Published Tests

In order to identify published criterion-referenced tests in reading, we contacted the Educational Testing Service and obtained from their Educational Test Collection Office lists of tests which were identified by their publishers as criterion-referenced tests. The tests were listed by levels as follows: preschool through grade 3, grades 4 through 6, and grade 7 and above. On the preschool through grade 3 level there were a total of about 200 entries, including separate entries by grade level for several achievement test batteries. Of the 200 entries, about 80 (or 40 percent) were either reading tests or reading readiness tests. The list of criterion-referenced tests for grades 4 through 6 contained a total of about 150 entries of which about 70 (or 45 percent) were reading tests. The list for grades 7 and above contained about 150 entries of which about one-third were reading tests. There was substantial overlap among the three lists.

We grouped the tests into the following categories:

1) Reading tests from achievement batteries. These are standardized tests which have been around for many years and which are now characterized as both norm-referenced and criterion-referenced.

2) Diagnostic tests (including reading readiness tests).

3) Reading Comprehension Tests.

4) Skills Monitoring System (reading management systems)

5) Survey Tests.

The comments which follow are addressed mainly to criterion-referenced tests called skills monitoring systems or reading management systems. Since the first of these was published in the early 1970's, their number has grown significantly. Indeed, it is in these systems that there has been the greatest proliferation of criterion-referenced tests in reading.

Skills monitoring systems are of two general types: program dependent and program independent. The former are interlocked with a single published instructional program, e.g. a basal reading series and are, in effect, a part of the program. Program independent systems are not integrated with any one published reading series, but may include charts which cross reference different basal reading series by page number.

These reading management systems are organized around objectives or subskills. The number of objectives varies from 60 to 70 in some systems to almost 400 in others. The objectives are sequenced in a hierarchical arrangement and the test items are keyed to each objective, the aim being to determine whether or not the student has acheived mastery of the objective. Since many of the tests have only two to three items, the student is frequently required to pass all items. Where there are more items, the passing score is commonly set arbitrarily at 80 or 90 percent. In some instances, however, the passing score varies depending upon the number of items on the test. In at least one system, no passing scores have been set; rather this is left to the discretion of the teacher or the local school system.

An important goal of reading management systems is to individualize instruction. Students who demonstrate mastery of an objective (or subskill) need not spend further time on it. On the other hand, for students who are

having difficulty in reading, the system should pinpoint their areas of difficulty and provide information for instructional decisions. One can hardly dispute the usefulness of such information.

The purpose of this paper is not to examine existing reading management systems in terms of their effectiveness in enabling the teacher to individualize instruction, but to look at the programs from the perspective of their reliability and validity. It should not be inferred that all systems are equivalent to each other in these regards. However, certain weaknesses which are common to these approaches to the evaluation of reading will be identified.

A notable weakness is the paucity of information reported on the reliability and validity of the tests used in these systems. This is a recurring theme in the test reviews found in Buros' Eighth Mental Measurements Yearbook (1968).

With respect to reliability, the brevity of most of the tests for specific objectives or subskills should again be noted. Commonly, there are two to three items, rarely more than five items, per objective. The items are of the multiple-choice type, generally with three to four alternatives. This probably would result in sizeable numbers of students being misclassified due to errors of measurement.

The validity issues are more complex. These issues include concerns about:

1) the view of reading and the process of learning to read which seems to be implicit in these reading management systems.
2) the hierarchical ordering of the objectives or subskills.
3) the lack of relevance of certain subskills which are identified as prerequisite to learning to read.
4) the mismatches between objectives and tests which purport to determine mastery of the objectives.

5) the arbitrary nature of cutting scores.

With regard to the first concern listed above, the perspective on the reading process implicit in the management systems, Johnson (See Buros, 1978) has noted that skills monitoring systems generally evoke a "reading-as-skill" perspective. Reading ability seems to be equated with the development of skills which can be used for instructional grouping. Learning to read is equated with mastery of these isolated skills. Thus, the reading act is fractionated into atomistic parts.

In addressing the second concern, the hierarchical ordering of skills, it should be noted that some sequencing of skills obviously is necessary in the teaching of reading -- one cannot teach everything at once. But a re-curring theme in reviews of these systems is that publishers imply that the particular sequencing in a given system represents the natural hierarchical order in which the acquisition of reading skills should occur.

The third concern relates to the irrelevance of certain subskills which are characterized as prerequisite to the acquisition of reading. Test publishers have provided little or no evidence that mastery of all the subskills included in instructional objectives are indeed necessary for reading. Some would seem to be irrelevant. For example, one system included the following motor skills:

Pupil is able to balance on one foot while teacher slowly counts to five.

Pupil is able to tie his shoe laces in a bow with ease.

Pupil is able to throw a beanbag into a wastepaper basket from a distance of five feet.

Pupil is able to catch a beanbag from a distance of eight feet.

The fourth concern, relating to the mismatch between items and objectives, really addresses the question of how well the test measures what it purports

to measure. Some examples from tests and comments upon them follow..

In one instance the teacher instructs first or second graders to "Find the word that has the same vowel sound as the underlined word." The children are assumed to know what is meant by the term "vowel sound." In another instance, four tests involving the matching of sounds may be completed on the basis of visual letter matching since the stimulus words to be spoken by the teacher are printed on the tests along with the response alternatives. Many of the so-called "phonics" tests are really spelling tests, involving such activities as listening to nonsense words and selecting a spelling that corresponds to a portion of the nonsense word.

With regard to the final concern, the validity of the mastery scores, it should be noted that passing scores are set arbitrarily. There is no evidence that students whose performance meets or exceeds the passing score have, in fact, mastered the subskill or objective or vice versa. Given the brevity of the tests, can one conceivably assess mastery? A further difficulty is inherent in the concept of mastery itself. With respect to word attack skills or word recognition it might be reasonable to talk about mastery at least in the narrow sense of the level of skill required to move on to a next level of reading instruction. However, as Johnson points out, the concept of mastery for reading subskills in tenous. Consider, for example, "Synonyms," a subtest at all levels of one of the testing systems. It does not seem possible to say that a child in the second grade can be said to have mastered synonyms in a generic sense, since not even Shakespeare managed that feat.

The foregoing comments may sound unduly negative, and it would certainly be unfair to paint all of the criterion-referenced tests in reading with the

same brush. The tests are not equal in quality and some may, in fact, be quite useful to the classroom teacher. This is a debated issue. But the tests do not meet the basic requirements of criterion-referenced testing within the classification schemes of Nitko, or Berk, or of Waldrop and his associates. The tests are, in Nitko's terms, referenced to ill-defined domains. They are, in Berk's terms, "mastery" tests, but not very good ones. Within the framework of Waldrop et al, the tests do not have a consistent inference system to support their intended usage.

Development of Domain-Reference Tests

There are no published domain-referenced tests of reading behavior. But there have been a number of experimental/research efforts to develop such tests. Bormuth's (1970) work should be mentioned here. Bormuth described an approach for objectively deriving test questions from prose content. In the Bormuth approach, test items are generated in two major steps: the analysis of the structure of prose content followed by the transformation of structure into test items. The goal is to generate test questions through the application of operationally defined rules, thus constraining the item-writer and eliminating or minimizing subjectivity in test development. Bormuth envisioned the generation of test items based upon discrete instructional sentences and also discourse-based items; i.e., items based upon relationships between sentences and groups of sentences.

Bormuth's book is not a compendium of rules and recipes, but it is considerably more definitive on some points than on others. With respect to "sentence verification" items and "wh-transformations," i.e., what, where, and who questions, the treatment is fairly concrete. It is less so in relation to what are labeled as "paraphrase transformations" and "semantic substitutes,"

and is still less definitive in dealing with the generation of discourse-based items.

Bormuth noted that the chief purpose of his book was to help provide a basis for the scientific study of acheivement test items. He identified a number of areas in which research was needed for the improvement of criterion-referenced testing, chief among which was the explication of the structure of instructional materials.

Cronbach (1970) reviewed Bormuth's book in Psychometrika. The review was not unmixed, but Cronbach found the ideas provocative and expressed the hope that the work might stimulate research that would be effective in changing test development from an experience-based judgemental process to a more systematic and scientific undertaking. Some of Bormuth's ideas, Cronbach observed, would take years of research to unravel but the effort might be worthwhile, especially the use of inter-sentence syntax and other means in combining elements hierarchically to derive test questions to measure higher mental processes.

Insofar as we are aware, Bormuth's book has not lead to such research, but it has lead to research of a more limited sort on the application of item-writing rules. We will not attempt to review these studies here. But two may be cited so as to suggest the nature of this research. Roid and Haladyna (1978) had two experienced item writers develop items on the same instructional materials, with each writer using two different techniques: generating items from objectives and from modified Bormuth-type rules. One writer consistently wrote more difficult items than the other, regardless of the item-writing technique. Both techniques resulted in about the same number of faulty items.

A study by Roid, Haladyna, and Shaughnessy (1980) contrasted three item-writing methods: informal, objective-based, and linguistic-based. The in-

formal method gave the greatest amount of freedom to the item writers. Linguistic methods imposed differing levels of constraint. In one case, item writers selected the foils. In another, foils were supplied by an algorithm. The informal method, the objective-based method, and the linguistic method which gave freedom in foil selection resulted in inter-writer differences in item difficulty and in item sensitivity to instruction. The linguistic method which supplied foils reduced these inter-writer differences, but resulted in easier items which were less sensitive to instruction. Foils supplied by algorithm tended to be less plausible than foils supplied by item writers.

Experimental Approaches

We wish now to describe briefly two studies which represent differing views on domain-referenced testing. The first is a study by Walmsley (1979). The study assessed the abilities of children to pronounce consonant-vowel-consonant (CVC) trigrams. There were a number of interesting findings, but our interest here is not so much in the findings and testing methods, as in the difficulties the author reported in defining the domain. It was a relatively simple domain which was selected partly because it seemed to lend itself to being fractionated, and thus clearly defined.

However, difficulties arose immediately in the definition of a consonant and a vowel. Is "r" a consonant? Can final "h," "w," and "y" be regarded as consonant letters, even though they do not represent consonant sounds in this position? Also, how does one differentiate between real words and nonsense syllables?

Walmsley felt that his resolution of such questions was arbitrary in the sense that another investigator might, with equal plausibility, arrive at different decisions in fractionating the domain. Certainly there are differing

views on the definition of reading. Walmsley observed that the choice of a particular universe of reading to be tested and the identification of its components reflect one's view of the definition and purpose of reading. The procedures followed in developing a criterion-referenced measure are useful in insuring that the items represent a particular universe, but they do not guarantee the validity of the universe itself.

Anderson et al. (1978) express a somewhat different perspective and describe a different approach to test development. Domain specification is held to be an interactive process involving interplay between conceptualization and observation. It is unrealistic to expect that a domain of instructional interest can be prespecified in all of its complexity. Thus one proceeds through conceptualization and observation to successively approximate the domain. Ineffect, one validates the domain.

As between the view that in domain-referenced testing one defines the domain and is then stuck with it and the view that domain specification and test development is interactive, we side with the latter. It seems to us that the results of Walmsley's study tend to support this interactive view, even though it seems at variance with the view he himself expresses. One of the dimensions on which Walmsley fractionated the CVC domain was the division between real words and nonsense syllables. The distinction was based upon entry in Websters International Dictionary (1972). This distinction was felt to be somewhat arbitrary. But in fact children did find CVC's which were real words easier to pronounce, even though they might be rare, obsolete, or slang words. Walmsley concluded, ". . .students in this sample appeared to make distinctions between CVC's that are real words (according to a strict dictionary definition) and nonsense syllables, despite the fact that some of

these real word CVC's must clearly have been outside the repertoires of most of the children in the sample." This finding seems to us to provide some degree of domain verification.

In concluding it is germane to comment on another point; namely, the level of effort required for the development of domain-referenced tests. Anderson et al. (1978) applied their approach experimentally to the development of what was in effect a "one item" test of reading comprehension and the effort was almost staggering. This level of effort may be justifiable for purposes of research in defining a domain, but it seems unlikely to result soon in domain-referenced tests for classroom use. What is needed, more realistically, are tests in between the so-called criterion-referenced tests reviewed in the Eighth Mental Measurements Yearbook and the domain-referenced tests described in the studies cited here; that is, tests which are feasible in terms of time and effort to develop, but which provide a sound informational base for instructional decisions.

## References


Anderson, T.H. and others, Developing and Trial of a Model for Developing Domain Referenced Tests of Reading Comprehension, Technical Report No. 86, Center for the Study of Reading, University of Illinois at Urbana-Champaign, 1978.

Berk, R.A. Domain-Referenced versus Mastery Conceptualization of Criterion-Referenced Measurement: A Clarification, Paper presented at the annual meeting of the American Educational Research Association, Boston, MA., April, 1980.

Bormuth, J.R. On the Theory of Achievement Test Items, Chicago: The University of Chicago Press, 1970.

Buros, O.K. "Fifty Years in Testing: Some Reminiscences, Criticisms and Suggestions." Educational Researchers, 1977, 6, 9-15.

Buros, O.K. (ed) The Eighth Mental Measurements Yearbook, Highland Park, N.J.: The Gryphon Press, 1978.

Cronback, J.L. John R. Bormuth: On the Theory of Achievement Test Items. Psychometrika, 1970, 35, 509-511. (Book Review).

Guttman, L. "Integration of Test Design and Analysis." In Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1969.

Jaeger, Richard, M. "The National Test-Equating study in Reading (The Anchor Test Study)" NCME Measurement in Education, vol. 4, No 4, Summer, 1973.

Nitko, A.J. "Distinguishing the Many Varieties of Criterion-Referenced Tests," Review of Educational Research, 1980, 50, 461-485.

Roid, G.H.and Haladyna, T.M. "A Comparison of Objective-Based and Modified Bormuth Item Writing Techniques," Educational and Psychological Measurement, 1978, 35, 19-28.

Roid, G., Haladyna, T., and Shaughnessy, J. A Comparison of Item-Writing Methods for Criterion-Referenced Tests, Paper presented at the annual meeting of the American Educational Research Association, Boston, MA., April, 1980.

Waldrop, J.L. and others. A framework for Analyzing Reading Test Characteristics, Technical Report No. 109, Center for the Study of Reading, University of Illinois at Urbana-Champaign, December, 1978.

Walmsley, S.A. "The Criterion-Referenced Measurement of an Early Reading Behavior," Reading Research Quarterly, XIV, 1979, 575-604.